**A tutorial for using the Matlab implementation of HMSC (Hierarchical Modelling of Species Communities)**

By Otso Ovaskainen, 14th September 2015

**General**

The Matlab version of HMSC consists of the main files A1\_generate\_data, A2\_HMSC, A3\_show\_estimates, and A4\_generate\_predictions. The main file A1 can be used to generate simulated data with known parameter values. The file A2 performs the posterior sampling. The file A3 generates some commonly needed summaries of the posterior distributions, and in case of simulated data compares them to true values. The file A4 can be used to generate model predictions based on the parameterized model. The other files are functions that are called from the main files A1-A4. The structural model definitions are set in the file model\_definitions. The priors are defined in the file define\_priors.

In addition to the beginnings of the files A1-A4 (i.e., the areas marked by comments), the user should modify only the functions model\_definitions and define\_priors as needed.

Put all HMSC functions to any directory that is included in the Matlab path.

**Part 1. Quick start with simulated data**

The quickest way to see how HMSC works is to generate simulated data and to estimate the parameters back from these data. Thus before starting to work with your own data, go through this example.

***Step 0. Create the folder structure and define the model structure***

Open model\_definitions and change the location of the basefolder from

basefolder='...\examples\';

to the path to the directory where you would like data and results from all HMSC analyses to be stored. Set the example to sim\_1 by letting

example = 'sim\_1';

and ensuring that the other options for the pre-made examples are commented away. This will define the model structure corresponding to simulate example number 1. In this simulated example, there are no traits, no phylogenetic constraints, nor spatially or temporally explicit structure. For a more detailed explanation of the choices related to model structure, see Part 2.

***Step 1. Generate simulated data***

Run the code in the file A1\_generate\_data (with example = ‘sim\_1’) to create simulated data. This will (through the call to model\_definitions) first create a working folder (inside the basefolder) called ‘sim\_1’, and within it four subfolders named ‘data’, ‘posteriors’, ‘results’ and ‘panels’ (unless these exist already). The simulated data will appear to the subfolder ‘data’. Explore the data files to understand their structure – when working with real data, you need to provide the data in the same format. With the example ‘sim\_1’, you will generate presence-absence data for a community with 100 sampling units, 20 species and 3 environmental covariates (including the intercept). The sampling units are assumed to belong to 10 plots.

The presence/absence data is in the file y\_sim.csv. The covariate data is in the file X\_sim.csv. The file LF\_units\_sim.csv describes what kind of random effects (modelled by latent variables) are included at each level. In this example, there are two columns. With the defaults values, the first column corresponds to a random effect at the sampling unit level, as each sampling unit has a unique number. The second column corresponds to a plot level effect, the number indicating which plot the sampling unit belongs to. The file dist\_sim.csv shows which statistical distribution is assumed – here the value is set to 2 for each species, corresponding to presence-absence data modelled with probit regression.

The files covariates\_sim.txt and species\_sim.txt include the names of the covariates and species (these two files are needed when plotting the results).

A subfolder ‘true values’ has been created inside the folder ‘data’. This contains the values of the parameters with which the data were created. These are stored so that the estimated values can be compared to the true values. Of course, with real data, the true values will not be known and thus this folder will not be present.

***Step 2. Estimate the parameters***

Run the code in the file A2\_HMSC to sample the posterior distributions. With the default settings (runs=5, samples=50, thinning=50; which you can modify in the beginning of this file), you will run the MCMC chain for 5 “runs”, each of which consists of 2500 iterations, from each of which 50 samples are stored. With the default settings (adapt=[2 0]), the numbers of the latent factors are adjusted during the first run, and then fixed for runs 2-5. The 5 “runs” are not parallel chains but all part of the same MCMC. The reason for having a number of “runs” is that with large datasets (and thus slow parameterization) the user may wish to set the code running for a long time (e.g. 100 runs), but check earlier (e.g. after 10 runs) if the chains have already mixed well enough.

The parameter values are stored in the subfolders ‘1\_sim’ … ‘5\_sim’, which should have appeared inside the folder “posteriors”.

***Step 3. Look at the results***

Run the code in the file A3\_show\_estimates to summarize and visualize the parameter estimates. In this file, you need to indicate which of the runs are to be included, which quantiles are to be included in the summaries, and if true values are available (i.e., if the data are simulate data). See the comments in the file A3 to see the meanings of these settings. Set the following to true or false according to which plots and files you wish to create

posterior\_summaries = true;

box\_plots = true;

mixing\_plots = true;

Generating the files and plots may take some time (e.g. 1 minute). Wait until the files are created before opening them.

The *posterior summaries* will appear to the folder “results”. The first sheet of each excel file shows the posterior means, the next ones the selected quantiles, and the last one the true values (if available). The species-specific regression parameters () are in the file beta\_sim.xlsx, the mean-vector () in the file gamma\_sim.xlsx. For the latent factor at level k=1,2, the variance-covariance matrices are in the file Omega\_k\_sim.xlsx. The corresponding correlations are in the file Rk\_sim.xlsx, and the partial correlations in the file iRk\_sim.xlsx.

MCMC *mixing plots* as well as some *box plots* (with subsets of species and covariates) will appear to the folder “panels”. These can be used to assess visually the convergence of the MCMC chains. If true values are available, they are included in the box plots as dots, and in the mixing plots as lines.

***Step 4. Make predictions***

The file A4\_generate\_predictions exemplifies how to make various kinds of predictions by HMSC. With example = ‘sim\_1’, it samples the parameters from the posterior, and then generates three kinds of posterior predictive data. First, it assumes the sampled latent factors, i.e. generating data for the same sampling units and plots that were present in the data used to estimate parameters. Second, it assumes unknown latent factors, i.e. utilizes the same fixed effects as when generating the data, but simulates the latent factors from their priors. Third, it samples the latent factors conditional on the knowledge of the occurrence of all other species than species 1-5.

**Part 2. General instructions**

Here we introduce the general settings of HMSC. In the following sections, they will be exemplified with case studies.

***Dimensions of the matrices***

Thorough, we use the following notation

* The number of species:
* The number of sampling units:
* The number of traits:
* The number of covariates:
* The number of latent factor effects (e.g. including “sampling unit level” and “plot level” would count as 2):

***Input parameters defined in the model\_definitions file***

* Set the working folder to the one you are using.
* Set “dataset” to correspond to the name of your data (the name should appear in all data files, see below).
* traits = true/false indicates if traits are to be included
* outlierspecies = true/false indicates if outlier species are assumed
* spatial = [true/false …. true/false] indicates if the latent factors for each level are spatial or not. The number of latent factor effects is deduced from this file. If spatial = [], there are no latent factors.
* Set to the number of latent factors

***Input files***

All input files are .csv files or .txt files (with ANSI encoding). All files are named so that they include the name of the dataset chosen by the user (in the above example, the name of the dataset is “\_sim”). The beginning of the filename tells which input file is in question.

*Obligatory input files*

* The file “dist\_sim.csv” is a matrix of values indicating which statistical distribution is assumed for the data for each species. This version of HMSC is restricted to dist=2 corresponding to probit regression (for binary data), other data models being under construction.
* The file “X\_sim.csv” is the matrix of the covariates .
* The file “y\_sim.csv” is the matrix of the data . Missing values are allowed for and are indicated as NaN. Missing values are not allowed for any other input files.
* The file “covariates\_sim.txt” includes the names of the covariates, each on its own row.
* The file “species\_sim.txt” includes the names of the species, each on its own row.

*Optional input files*

* If traits are included…
  + The file “T\_sim.csv” is the matrix of the traits
  + The file “traits\_sim.txt” includes the names of the traits, each on its own row.
* If latent factors are included, the file “LF\_units\_sim.csv” it is a matrix, where the entries give the level to which the sampling unit belongs to. For example, for a plot level effect, the levels are the identity numbers of the plots.
* If latent factor at level is spatial…
  + The file “LF\_alpha\_sim\_i.csv” gives the discrete grid prior for alpha. The first column gives the values, the second their weights.
  + The file “LF\_xy\_sim\_i.csv” gives the spatial (or temporal) coordinates of the units. The number of columns corresponds to the spatial dimension and can be arbitrary.

**Part 3. Example data sets**

In all examples, we give below some context, describe the data, and indicate what settings should be applied to run the parameter estimation. Once the parameter estimation has finished, run the file A3\_show\_estimates to produce the mixing plots and some posterior summaries. These appear to the folders “results” and “panels”. You may then run the file A4\_generate\_predictions, e.g. to examine model fit by comparing posterior predictive data to the original data or (with some of the examples) to independent validation data.

**Simulated data set 1**

In files model\_definitions, A1\_generate\_data and A4\_generate\_predictions, set example = ‘sim\_1’. In this simulated example, there are no traits nor spatially or temporally explicit structure. The example involves presence-absence data for a community with 100 sampling units, 20 species and 3 environmental covariates (including the intercept). The sampling units are assumed to belong to 10 plots.

**Spatial model with Great Britain butterfly data**

Here we replicate the analyses of the paper “Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models” by Ovaskainen et al. (Methods in Ecology and Evolution, in press). There are 55 species, 3 traits, and 5 covariates (see the paper and the files to see what these are).

In files model\_definitions and A4\_generate\_predictions, set example = ‘butterfly’. The data are found from the folder ‘butterfly’. The extension ‘WLF\_300’ means that the model indicates latent factors (WLF = with latent factors) and that the data are split to 300 training locations and the remaining validation locations. The extension ‘WOLF\_300’ means that the model does not include latent factors (WLF = without latent factors). In the file model\_definitions, choose if you wish to include or exclude the latent factors by commenting away either dataset=’\_WLF\_300’ and spatial=[true], or dataset=’\_WOLF\_300’ and spatial=[]. There are two copies of each data file, one with the extension “v” (standing for validation), and one without that extension (the training data).

The number of latent factors can either be fixed or to be estimated. In the paper, the number of latent factors was estimated in the initial run. The results from that indicated that only two factors made a negligible contribution, so to simplify the analyses, the final results were obtained by fixing the number of factors to two. To keep it fixed, in the file 2\_HMSC.m set fix\_nf = true, and set the initial value to the desired value in the file compute\_initial values (in the paper, this was done by setting nf = ones(nr,1)\*2).

This is computationally somewhat challenging problem. Running the MCMC for runs=5, samples=100, thinning=100 is likely to take ca. 2 hrs.

The results of the paper can be generated by running the additional file ‘A5\_butterfly’. This file samples the posterior, generates for each sample a variance partitioning, the expected latent factors for the validation sites, and a simulated species community for the training and validation sites. The results are averaged over 100 samples from the posterior and stored into a folder called ‘predictions’.